

# CONCENTRATION BOUNDS FOR ENTROPY ESTIMATION OF ONE-DIMENSIONAL GIBBS MEASURES

J.-R. CHAZOTTES, C. MALDONADO

**ABSTRACT.** We obtain bounds on fluctuations of two entropy estimators for a class of one-dimensional Gibbs measures on the full shift. They are the consequence of a general exponential inequality for Lipschitz functions of  $n$  variables. The first estimator is based on empirical frequencies of blocks scaling logarithmically with the sample length. The second one is based on the first appearance of blocks within typical samples.

## CONTENTS

1. Introduction	1
2. Setting	2
2.1. Notations and definitions	2
2.2. Entropy	3
2.3. Gibbs measures	3
3. An exponential inequality and its general consequences	3
3.1. An exponential inequality	3
3.2. General consequences	4
4. Bounds on entropy estimators	5
4.1. Plug-in estimator	6
4.2. Hitting times	8
Appendix A. Proof of Lemma 4.1	11
References	12

## 1. INTRODUCTION

Given a ‘sample’  $x_0, x_1, \dots$  of a finite-valued discrete-time ergodic process  $\{X_n; n \in \mathbb{N}\}$ , there are several ways to consistently estimate its entropy. In this paper we shall study two estimators. One is based on empirical frequencies of blocks and is referred to as the ‘plug-in’ estimator. The other one is based on the first appearance or repetition of blocks within the sample. We refer to [13] for their basic properties. Here we are concerned with the fluctuation properties of these estimators. We will further assume that the joint distribution of the process  $\{X_n; n \in \mathbb{N}\}$  is Gibbsian in a way made precise below.

Fluctuations of the plug-in estimator were already studied in [10] and [5] from the viewpoint of classical limit theorems. Namely, in [10] the authors prove a central limit theorem and in [5] a large deviation principle is obtained.

Regarding the return-time and the hitting-time estimators, previous results are found in [7] and [6]. Central limit theorems and large deviations principle are established in these papers. In the present article, we only study the hitting-time estimator.

Our aim is to obtain bounds on the fluctuations of the plug-in and hitting-time estimator in the spirit of concentration inequalities. Concentration inequalities became recently a widespread powerful tool in many fields of pure and applied probabilities, as well as in functional analysis, combinatorics, computer science, etc; see for instance [11] and [9]. In the context of dynamical systems, the first result was proved in [8] where several applications are presented (see also [4]). Namely, an exponential inequality is proved for any separately Lipschitz function of  $n$  variables for a class of piecewise expanding maps of the interval. In our setting the same inequality holds. The proof is the same as in [8]. It is in fact simpler since no Markov partition is assumed therein. In this paper we apply this exponential inequality to get some fluctuation bounds on our entropy estimators. The only previous work where this is done for the plug-in estimator is found in [2] in the case where the  $X_i$ 's are independent identically distributed random variables taking values in a countable set. For the hitting-time estimator, no such bounds were known before, even in the case of independent random variables.

Our main results are theorems 4.1, 4.2 and 4.3. Let us emphasize that we establish bounds for every  $n$ ,  $n$  being the sample length, whereas the results obtained in [5, 6, 7, 10] are in some sense finer but they are only asymptotic. Therefore, our work complement the picture on the fluctuations of these entropy estimators. Theorems 4.1 and 4.2 concern the plug-in estimator. Theorem 4.3 is about the hitting-time estimator. It should be noted that the route to get the bounds is not as direct as for the plug-in estimator because the hitting-time a priori behaves badly. The trick is to take advantage of its approximation by the inverse measure of the corresponding cylinder. This is where Gibbsianness is crucial.

This paper is organized as follows. In Section 2 we recall some definitions and facts, and state the exponential bound from which concentration inequalities follow. Section 4 contains our results on the plug-in estimators and the hitting-time estimator.

## 2. SETTING

After fixing some notations, we recall a few facts about entropy and Gibbs measures.

**2.1. Notations and definitions.** We consider the set  $\Omega = A^{\mathbb{N}}$  of infinite sequences  $\underline{x}$  of symbols from the finite set  $A$ :  $\underline{x} = x_0, x_1, \dots$  where  $x_j \in A$ . We denote by  $\sigma$  the shift map on  $\Omega$ :  $(\sigma \underline{x})_i = x_{i+1}$ , for all  $i = 0, 1, \dots$

We equip  $\Omega$  with the usual distance: fix  $\theta \in (0, 1)$  and for  $\underline{x} \neq \underline{y}$ , let  $d_\theta(\underline{x}, \underline{y}) = \theta^N$  where  $N$  is the largest nonnegative integer with  $x_i = y_i$  for every  $0 \leq i < N$ . (By convention, if  $\underline{x} = \underline{y}$  then  $N = \infty$  and  $\theta^\infty = 0$ , while if  $x_0 \neq y_0$  then  $N = 0$ .) With this distance  $\Omega$  is a compact metric space.

For a given string  $a_0^{k-1} = a_0, \dots, a_{k-1}$  ( $a_i \in A$ ), the set  $[a_0^{k-1}] = \{\underline{x} \in \Omega : x_i = a_i, i = 0, \dots, k-1\}$  is the cylinder of length  $k$  based on  $a_0, \dots, a_{k-1}$ .

For a continuous function  $f : \Omega \rightarrow \mathbb{R}$  and  $m \geq 0$  we define

$$\text{var}_m(f) := \sup\{|f(\underline{x}) - f(\underline{y})| : x_i = y_i, i = 0, \dots, m\}.$$

It is easy to see that  $|f(\underline{x}) - f(\underline{y})| \leq C d_\theta(\underline{x}, \underline{y})$  if and only if  $\text{var}_m(f) \leq C \theta^m$ ,  $m = 0, 1, \dots$ . Let

$$\mathcal{F}_\theta = \{f : f \text{ continuous, } \text{var}_m(f) \leq C \theta^m, m = 0, 1, \dots, \text{ for some } C > 0\}.$$

This is the space of Lipschitz functions with respect to the distance  $d_\theta$ . For  $f \in \mathcal{F}_\theta$  let  $|f|_\theta = \sup \left\{ \frac{\text{var}_m(f)}{\theta^m} : m \geq 0 \right\}$ . We notice that  $|f|_\theta$  is merely the least Lipschitz constant of  $f$ .

**2.2. Entropy.** Let  $\nu$  be a shift-invariant probability measure on  $\Omega$  and

$$H_k(\nu) = - \sum_{a_0^{k-1} \in A^k} \nu([a_0^{k-1}]) \log \nu([a_0^{k-1}]),$$

its ‘ $k$ -block entropy’. Then the entropy of  $\nu$  is

$$h(\nu) = \lim_{k \rightarrow \infty} \frac{H_k(\nu)}{k}.$$

Recall that  $0 \leq h(\nu) \leq \log |A|$ .

**2.3. Gibbs measures.** Full details for this section can be found in [3]. Let  $\phi \in \mathcal{F}_\theta$  and  $\mu_\phi$  the associated Gibbs measure. It is the unique shift-invariant probability measure for which one can find constants  $C = C(\phi) > 1$  and  $P = P(\phi)$  such that

$$(1) \quad C^{-1} \leq \frac{\mu_\phi(\{y : y_i = x_i, \forall i \in [0, m)\})}{\exp\left(-Pm + \sum_{k=0}^{m-1} \phi(\sigma^k \underline{x})\right)} \leq C$$

for every  $\underline{x} \in \Omega$  and  $m \geq 1$ . The constant  $P$  is the topological pressure of  $\phi$ . We can always assume that  $P = 0$  by considering the potential  $\phi - P$  which yields the same Gibbs measure.

The Gibbs measure  $\mu_\phi$  satisfies the variational principle, namely

$$\sup \left\{ h(\eta) + \int \phi d\eta : \eta \text{ shift-invariant} \right\} = h(\mu_\phi) + \int \phi d\mu_\phi = P = 0.$$

More precisely,  $\mu_\phi$  is the unique shift-invariant measure reaching this supremum. In particular we have

$$(2) \quad h(\mu_\phi) = - \int \phi d\mu_\phi.$$

### 3. AN EXPONENTIAL INEQUALITY AND ITS GENERAL CONSEQUENCES

**3.1. An exponential inequality.** Our main tool is an exponential inequality for fairly general observables.

Let  $K : \Omega^n \rightarrow \mathbb{R}$  be a function of  $n$  variables and, for each  $j = 0, \dots, n-1$ , let

$$\text{Lip}_j(K) = \sup_{\underline{x}^{(0)}, \underline{x}^{(1)}, \dots, \underline{x}^{(n-1)}} \sup_{\underline{y}^{(j)} \neq \underline{x}^{(j)}} \frac{|K(\underline{x}^{(0)}, \dots, \underline{x}^{(j-1)}, \underline{x}^{(j)}, \underline{x}^{(j+1)}, \dots, \underline{x}^{(n-1)}) - K(\underline{x}^{(0)}, \dots, \underline{x}^{(j-1)}, \underline{y}^{(j)}, \underline{x}^{(j+1)}, \dots, \underline{x}^{(n-1)})|}{d_\theta(\underline{x}^{(j)}, \underline{y}^{(j)})}.$$

We shall say that  $K$  is a separately Lipschitz function of  $n$  variables if

$$\text{Lip}_j(K) < \infty, \quad j = 0, \dots, n-1.$$

We now present our main tool.

**Theorem 3.1** ([8]). *Let  $\mu_\phi$  be a Gibbs measure. Then there exists a constant  $D = D(\phi) > 0$  such that, for any integer  $n \geq 1$  and for any separately Lipschitz function  $K$  of  $n$  variables, one has*

$$(3) \quad \int e^{K(\underline{x}, \dots, \sigma^{n-1} \underline{x})} d\mu_\phi(\underline{x}) \leq e^{\int K(\underline{y}, \dots, \sigma^{n-1} \underline{y}) d\mu_\phi(\underline{y})} e^{D \sum_{i=0}^{n-1} \text{Lip}_i^2(K)}.$$

Let us emphasize that the constant  $D$  only depends on  $\phi$ . It depends neither on  $K$  nor on  $n$ .

The powerfulness of (3) lies in that it applies to *any* separately Lipschitz function of  $n$  variables, regardless of its complicated or implicit form. All we have to do is to estimate its Lipschitz constants.

The proof of Theorem (3.1) easily follows from [8] where it is done for a class of piecewise expanding maps of the interval (without assuming a Markov partition). In fact, the proof becomes simpler in our setting. It relies on the fact that the transfer operator associated to  $\phi$  has a spectral gap when acting on Lipschitz functions. The point is to write an observable of  $n$  variables as a telescopic sum of observables depending only on one variable.

**3.2. General consequences.** We derive several consequences of inequality (3). The first one is a bound for the probability of  $K$  to deviate from its expectation.

**Corollary 3.1.** *For every  $t > 0$ , one has*

$$(4) \quad \mu_\phi \left\{ \underline{x} : K(\underline{x}, \dots, \sigma^{n-1} \underline{x}) \geq \int K(\underline{y}, \dots, \sigma^{n-1} \underline{y}) d\mu_\phi(\underline{y}) + t \right\} \leq e^{-\frac{t^2}{4D \sum_{i=0}^{n-1} \text{Lip}_i^2(K)}}.$$

*Proof.* The proof is an immediate consequence of Markov inequality and (3): for every  $\lambda > 0$ , the function  $\lambda K$  is separately Lipschitz and

$$\begin{aligned} & \mu_\phi \left\{ \underline{x} : K(\underline{x}, \dots, \sigma^{n-1} \underline{x}) \geq \int K(\underline{y}, \dots, \sigma^{n-1} \underline{y}) d\mu_\phi(\underline{y}) + t \right\} \\ & \leq e^{-\lambda t} \int e^{\lambda [K(\underline{x}, \dots, \sigma^{n-1} \underline{x}) - \int K(\underline{y}, \dots, \sigma^{n-1} \underline{y}) d\mu_\phi(\underline{y})]} d\mu_\phi(\underline{x}) \\ & \leq e^{-\lambda t + \lambda^2 D \sum_{i=0}^{n-1} \text{Lip}_i^2(K)}. \end{aligned}$$

It remains to optimize over  $\lambda$  to get the desired inequality. □

Of course we can apply (4) to  $-K$  and get by a union bound that

$$\mu_\phi \left\{ \underline{x} : \left| K(\underline{x}, \dots, \sigma^{n-1} \underline{x}) - \int K(\underline{y}, \dots, \sigma^{n-1} \underline{y}) d\mu_\phi(\underline{y}) \right| \geq t \right\} \leq 2 e^{-\frac{t^2}{4D \sum_{i=0}^{n-1} \text{Lip}_i^2(K)}}$$

for every  $t > 0$ .

Another immediate consequence of (3) is a bound on the variance of  $K$ :

**Corollary 3.2.** *One has*

$$\int \left( K(\underline{x}, \dots, \sigma^{n-1}\underline{x}) - \int K(\underline{y}, \dots, \sigma^{n-1}\underline{y}) \, d\mu_\phi(\underline{y}) \right)^2 d\mu_\phi(\underline{x}) \leq 2D \sum_{i=0}^{n-1} \text{Lip}_i^2(K).$$

*Proof.* We apply (3) to  $\lambda K$ , with  $\lambda \neq 0$ , to get at once

$$\frac{1}{\lambda^2} \left( \int e^{\lambda [K(\underline{x}, \dots, \sigma^{n-1}\underline{x}) - \int K(\underline{y}, \dots, \sigma^{n-1}\underline{y}) \, d\mu_\phi(\underline{y})]} d\mu_\phi(\underline{x}) - 1 \right) \leq \frac{1}{\lambda^2} \left( e^{\lambda^2 D \sum_{i=0}^{n-1} \text{Lip}_i^2(K)} - 1 \right).$$

The result follows by Taylor expansion and letting  $\lambda$  going to 0.  $\square$

The simplest, yet non-trivial, application of the above results is to ergodic sums, that is to take  $K_0(\underline{x}^{(0)}, \underline{x}^{(1)}, \dots, \underline{x}^{(n-1)}) = f(\underline{x}^{(0)}) + f(\underline{x}^{(1)}) + \dots + f(\underline{x}^{(n-1)})$  where  $f : \Omega \rightarrow \mathbb{R}$  is Lipschitz. A particular case of Corollary 3.1 yields immediately the following result, stated for later convenience.

**Corollary 3.3.** *Let  $f : \Omega \rightarrow \mathbb{R}$  be a Lipschitz function. Then*

$$(5) \quad \mu_\phi \left\{ \underline{x} : \frac{1}{n} (f(\underline{x}) + \dots + f(\sigma^{n-1}\underline{x})) - \int f d\mu_\phi \geq t \right\} \leq e^{-Bnt^2}$$

for every  $t > 0$  and for every  $n \geq 1$ , where  $B := (4D|f|_\theta^2)^{-1}$ .

We can of course apply (5) to  $-f$  to get

$$(6) \quad \mu_\phi \left\{ \underline{x} : \frac{1}{n} (f(\underline{x}) + \dots + f(\sigma^{n-1}\underline{x})) - \int f d\mu_\phi \leq -t \right\} \leq e^{-Bnt^2}.$$

By a union bound, (5) and (6) yield at once

$$\mu_\phi \left\{ \underline{x} : \left| \frac{1}{n} (f(\underline{x}) + \dots + f(\sigma^{n-1}\underline{x})) - \int f d\mu_\phi \right| \geq t \right\} \leq 2e^{-Bnt^2}.$$

In words, the ergodic average of  $f$  concentrates sharply around its  $\mu_\phi$ -average. The bound is exponentially small in  $n$  and when  $t$  gets large, the probability of deviation is extremely small.

Let us close this section by a basic observation. Many estimators of interest are functions of  $n$  symbols, that is, functions of the form  $\tilde{K} : A^n \rightarrow \mathbb{R}$ . A function  $\tilde{K} : A^n \rightarrow \mathbb{R}$  can be identified with a function  $K : \Omega^n \rightarrow \mathbb{R}$ . When applying Theorem 3.1 and its corollaries in this special case,  $\text{Lip}_j(K)$  has to be replaced by  $\delta_j(\tilde{K})$ , the oscillation at the  $j$ -th coordinate, where

$$(7) \quad \delta_j(\tilde{K}) = \sup_{a_0, \dots, a_{n-1}} \sup_{a_j \neq b_j} \left| \tilde{K}(a_0, \dots, a_{j-1}, a_j, a_{j+1}, \dots, a_{n-1}) - \tilde{K}(a_0, \dots, a_{j-1}, b_j, a_{j+1}, \dots, a_{n-1}) \right|.$$

#### 4. BOUNDS ON ENTROPY ESTIMATORS

Throughout this section,  $\phi \in \mathcal{F}_\theta$  and  $\mu_\phi$  is its unique Gibbs measure.

**4.1. Plug-in estimator.** The plug-in estimator is based on the empirical frequency of a word  $a_0^{k-1}$  in a ‘sample’  $x_0, x_1, \dots, x_{n-1}$ :

$$\mathcal{E}_k(a_0^{k-1}; x_0^{n-1}) = \frac{1}{n} \# \{0 \leq j \leq n : \tilde{x}_j^{j+k-1} = a_0^{k-1}\},$$

where  $\tilde{x} := x_0^{n-1} x_0^{n-1} \dots$  is the periodic point with period  $n$  made from  $x_0^{n-1}$ . This trick makes  $\mathcal{E}_k(\cdot; x_0^{n-1})$  a locally shift-invariant probability measure on  $A^k$ .

For any ergodic measure  $\nu$ , there is a set of  $\nu$ -measure one of  $\underline{x}$ ’s such that for every  $k \geq 1$

$$\lim_{n \rightarrow \infty} \mathcal{E}_k(a_0^{k-1}; x_0^{n-1}) = \nu([a_0^{k-1}]).$$

The  $k$ -block empirical entropy is defined as

$$\hat{H}_k(x_0^{n-1}) := H_k(\mathcal{E}_k(\cdot; x_0^{n-1})).$$

It is clear that for  $\nu$ -almost every  $\underline{x}$

$$\lim_{k \rightarrow \infty} \lim_{n \rightarrow \infty} \frac{\hat{H}_k(x_0^{n-1})}{k} = h(\nu).$$

As shown by Ornstein and Weiss (see [13]), we can in fact take a single limit by letting  $k$  depend on  $n$ : if  $k(n) \rightarrow \infty$  and  $k(n) \leq \frac{1}{h(\nu)} \log n$  then

$$\lim_{n \rightarrow \infty} \frac{\hat{H}_{k(n)}(x_0^{n-1})}{k(n)} = h(\nu) \quad \text{for } \nu - \text{almost-every } \underline{x}.$$

Note that since  $h(\nu) \leq \log |A|$  we can always take  $k(n) \leq \frac{1}{\log |A|} \log n$ .

We can formulate our first result on fluctuations of the plug-in entropy estimator. We denote by  $\mathbb{E}$  the expectation and by  $\text{Var}$  the variance under  $\mu_\phi$ .

**Theorem 4.1.** *Let  $D$  be the constant appearing in (3). For every  $\alpha \in (0, 1)$ ,  $t > 0$  and  $n \geq 2$  one has*

$$\mu_\phi \left\{ \left| \frac{\hat{H}_{k(n)}}{k(n)} - \mathbb{E} \left( \frac{\hat{H}_{k(n)}}{k(n)} \right) \right| \geq t \right\} \leq 2 \exp \left( - \frac{n^{1-\alpha} t^2}{16D(\log n)^2} \right)$$

provided that  $k(n) \leq \frac{\alpha}{2 \log |A|} \log n$ .

Moreover for every  $n \geq 2$

$$\text{Var} \left( \frac{\hat{H}_{k(n)}}{k(n)} \right) \leq 8D \frac{(\log n)^2}{n^{1-\alpha}}.$$

*Proof.* Given any integer  $k \geq 1$ , consider the function  $\tilde{K} : A^n \rightarrow \mathbb{R}$  defined by

$$\tilde{K}(s_0, \dots, s_{n-1}) = \hat{H}_k(s_0^{n-1}).$$

We estimate the  $\delta_j(\tilde{K})$ ’s (see (7) for the definition of  $\delta_j(\cdot)$ ). We claim that

$$\delta_j(\tilde{K}) \leq 2k|A|^k \frac{\log n}{n}, \quad \forall j = 0, \dots, n-1.$$

Indeed, given any string  $a_0^{k-1}$ , the change of one symbol in  $s_0^{n-1}$  can decrease  $\mathcal{E}(a_0^{k-1}; s_0^{n-1})$  by at most  $k/n$ . It is possible that another string gets its frequency increased, and this

can be at most by  $k/n$ . This is the worst case. We then use the fact that for any pair of positive integers  $l$  and  $k$  such that  $l + k \leq n$ , one has

$$\left| \left( \frac{l}{n} \right) \log \left( \frac{l}{n} \right) - \left( \frac{l+k}{n} \right) \log \left( \frac{l+k}{n} \right) \right| \leq \frac{k}{n} \log n.$$

The claim follows by summing up this bound for all strings, which gives the factor  $|A|^k$ . Finally, taking  $k(n) \leq \frac{\alpha}{2 \log |A|} \log n$ , with  $\alpha \in (0, 1)$ , and applying Corollaries 3.1 and 3.2, we get the desired bounds.  $\square$

It is natural to seek for a concentration bound for the empirical entropy not about its expectation, but about  $h(\mu_\phi)$ , the entropy of the Gibbs measure. To have good control on this expectation, it turns out that a better estimator is the conditional empirical entropy. To define it, we need to recall a few definitions and facts.

For a shift-invariant measure  $\nu$  and  $k \geq 2$ , let

$$h_k(\nu) = H_k(\nu) - H_{k-1}(\nu) = - \sum_{a_0^{k-1}} \nu([a_0^{k-1}]) \log \frac{\nu([a_0^{k-1}])}{\nu([a_0^{k-2}])}.$$

It is well-known that  $\lim_{k \rightarrow \infty} h_k(\nu) = h(\nu)$  (see for instance [13]).

The  $k$ -block conditional empirical entropy is

$$\hat{h}_k(x_0^{n-1}) = \hat{h}_k(\mathcal{E}_k(\cdot; x_0^{n-1})).$$

When  $\nu$  is ergodic, one can prove [13] that, if  $k(n) \rightarrow \infty$  and  $k(n) \leq \frac{(1-\epsilon)}{\log |A|} \log n$ , for any  $\epsilon \in (0, 1)$ , then

$$\lim_{n \rightarrow \infty} \hat{h}_{k(n)}(x_0^{n-1}) = h(\nu), \quad \text{for } \nu - \text{almost every } \underline{x}.$$

We have the following result.

**Theorem 4.2.** *Assume that  $\theta < |A|^{-1}$ . There exist strictly positive constants  $c, \gamma, \Gamma, \xi$  such that for every  $t > 0$  and for every  $n$  large enough*

$$\mu_\phi \left\{ \left| \hat{h}_{k(n)} - h(\mu_\phi) \right| \geq t + \frac{c}{n^\gamma} \right\} \leq 2 \exp \left( - \frac{\Gamma n^\xi t^2}{(\log n)^4} \right)$$

*provided that  $k(n) < \frac{\log n}{2 \log |A|}$ .*

**Remark 4.1.** *From the proof we have  $\gamma = 1/(1 + \frac{\log |A|}{\log(\theta^{-1})})$ ,  $\xi = 1 - 2/(1 + \frac{\log(\theta^{-1})}{\log |A|})$  and  $\Gamma = (\log |A|)^2/16D$ .*

*Proof.* By definition  $\hat{h}_k = \hat{H}_k - \hat{H}_{k-1}$ . If we let  $\tilde{K}'(s_0, \dots, s_{n-1}) = \hat{h}_k(s_0^{n-1})$ , we estimate  $\delta_j(\tilde{K}')$  by  $2\delta_j(\tilde{K})$ .

We now estimate the expectation of  $\hat{h}_{k(n)}$ . We need the following lemma.

**Lemma 4.1.** *We have*

$$(8) \quad \hat{h}_{k(n)}(x_0^{n-1}) = \frac{1}{n} \sum_{j=0}^{n-1} (-\phi(\sigma^j \underline{x})) + \hat{\Delta}_{k(n)}(x_0^{n-1}) + \mathcal{O}(\theta^{k(n)})$$

where

$$(9) \quad |\mathbb{E}(\hat{\Delta}_{k(n)})| \leq \frac{M|A|^{k(n)}}{n},$$

where  $M > 0$ .

This lemma can be deduced from the proof of Theorem 2.1 in [10]. However, for the reader's convenience, we provide part of its proof in the appendix.

Now subtract  $h(\mu_\phi)$  and take the expectation on both sides of (8), to get, using (2),

$$\mathbb{E}(\hat{h}_{k(n)}) - h(\mu_\phi) = \mathbb{E}(\hat{\Delta}_{k(n)}) + \mathcal{O}(\theta^{k(n)}).$$

We now take  $k(n) = q \log n / \log |A|$ , where  $0 < q < 1$  has to be determined. Choosing  $q = 1/(1 + \frac{\log \theta^{-1}}{\log |A|})$  we easily get that

$$(10) \quad |\mathbb{E}(\hat{h}_{k(n)}) - h(\mu_\phi)| \leq \frac{c}{n^\gamma},$$

where  $c > 0$  is some constant and  $\gamma = 1/(1 + \frac{\log |A|}{\log(\theta^{-1})})$ .

To end the proof, we apply Corollary 3.1 and use (10). For the exponent  $\xi$  in the statement of the theorem be strictly positive, one must have  $q < 1/2$ , which is equivalent to the requirement that  $\theta < |A|^{-1}$ .  $\square$

**4.2. Hitting times.** Given  $\underline{x}, \underline{y} \in \Omega$ , let

$$W_n(\underline{x}, \underline{y}) = \inf\{j \geq 1 : y_j^{j+n-1} = x_0^{n-1}\}.$$

Under suitable mixing conditions on the shift-invariant measure  $\nu$ , one can prove [13] that

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log W_n(\underline{x}, \underline{y}) = h(\nu), \quad \text{for } \nu \otimes \nu - \text{almost every } (\underline{x}, \underline{y}).$$

In particular, when  $\nu$  is a Gibbs measure in the above sense, this result holds true [6].

We have the following concentration bounds for the hitting-time estimator.

**Theorem 4.3.** *There exist constants  $C_1, C_2 > 0$  and  $t_0 > 0$  such that, for every  $n \geq 1$  and every  $t > t_0$ ,*

$$(11) \quad (\mu_\phi \otimes \mu_\phi) \left\{ (\underline{x}, \underline{y}) : \frac{1}{n} \log W_n(\underline{x}, \underline{y}) > h(\mu_\phi) + t \right\} \leq C_1 e^{-C_2 n t^2}$$

and

$$(12) \quad (\mu_\phi \otimes \mu_\phi) \left\{ (\underline{x}, \underline{y}) : \frac{1}{n} \log W_n(\underline{x}, \underline{y}) < h(\mu_\phi) - t \right\} \leq C_1 e^{-C_2 n t}.$$

Let us notice that the upper tail estimate behaves differently than the lower tail estimate as a function of  $t$ . This asymmetric behavior also shows up in the large deviation asymptotics [6].

Let us sketch the strategy to prove Theorem 4.3. We cannot apply directly our concentration inequality to the random variable  $W_n$  for the following basic reason. Given  $\underline{x}$  and  $\underline{y}$ , the first time that one sees the first  $n$  symbols of  $\underline{x}$  in  $\underline{y}$  is  $W_n(\underline{x}, \underline{y})$  and assume it is finite. If we make  $\underline{y}'$  by changing one symbol in  $\underline{y}$ , we have a priori no control on  $W_n(\underline{x}, \underline{y}')$  which can be arbitrarily larger than  $W_n(\underline{x}, \underline{y})$  and even infinite. Of



course, this situation is not typical, but we are forced to use the worst case to apply our concentration inequality. Roughly, we proceed as follows. We obviously have  $\log W_n = \log(W_n \mu_\phi([X_0^{n-1}])) - \log \mu_\phi([X_0^{n-1}])$ . On the one hand, we use a sharp approximation of the law of the random variables  $W_n \mu_\phi([X_0^{n-1}])$  by an exponential law proved in [1]. On the other hand, by the Gibbs property,  $\log \mu_\phi([x_0^{n-1}]) \approx \phi(\underline{x}) + \dots + \phi(\sigma^{n-1} \underline{x})$  and we can use Corollary 3.3 for  $f = \phi$ .

*Proof.* We first prove (11). We obviously have

$$\begin{aligned} & (\mu_\phi \otimes \mu_\phi) \left\{ (\underline{x}, \underline{y}) : \frac{1}{n} \log W_n(\underline{x}, \underline{y}) > h(\mu_\phi) + t \right\} \\ &= (\mu_\phi \otimes \mu_\phi) \left\{ (\underline{x}, \underline{y}) : \frac{1}{n} \log W_n(\underline{x}, \underline{y}) + \frac{1}{n} \log \mu_\phi([x_0^{n-1}]) - \frac{1}{n} \log \mu_\phi([x_0^{n-1}]) - h(\mu_\phi) > t \right\} \\ &\leq (\mu_\phi \otimes \mu_\phi) \left\{ (\underline{x}, \underline{y}) : \frac{1}{n} \log [W_n(\underline{x}, \underline{y}) \mu_\phi([x_0^{n-1}])] > \frac{t}{2} \right\} \\ &\quad + \mu_\phi \left\{ \underline{x} : -\frac{1}{n} \log \mu_\phi([x_0^{n-1}]) - h(\mu_\phi) > \frac{t}{2} \right\} \\ &=: T_1 + T_2. \end{aligned}$$

We first derive an upper bound for  $T_2$ .

We use (1), Corollary 3.3 applied to  $f = -\phi$  and (2) to get

$$\begin{aligned} T_2 &\leq \mu_\phi \left\{ -\frac{1}{n} (\phi + \dots + \phi \circ \sigma^{n-1}) - h(\mu_\phi) > \frac{t}{2} - \frac{1}{n} \log C \right\} \\ &\leq e^{-Bnt^2} \end{aligned}$$

for every  $t$  larger than  $2 \log C$ .

We now derive an upper bound for  $T_1$ . To this end we apply the following result which we state as a lemma. It is an immediate consequence of Theorem 1 in [1].

**Lemma 4.2** ([1]). *Let*

$$\tau_{[a_0^{n-1}]}(\underline{y}) := \inf \{j \geq 1 : y_j^{j+n-1} = a_0^{n-1}\}.$$

*There exist strictly positive constants  $C, c, \lambda_1, \lambda_2$ , with  $\lambda_1 < \lambda_2$ , such that for every  $n \in \mathbb{N}$ , every string  $a_0^{n-1}$ , there exists  $\lambda(a_0^{n-1}) \in [\lambda_1, \lambda_2]$  such that*

$$\left| \mu_\phi \left\{ \underline{y} : \tau_{[a_0^{n-1}]}(\underline{y}) > \frac{u}{\lambda(a_0^{n-1}) \mu_\phi([a_0^{n-1}])} \right\} - e^{-u} \right| \leq C e^{-cu}$$

*for every  $u > 0$ .*

By definition and using the previous lemma we get

$$\begin{aligned} T_1 &= \sum_{a_0^{n-1}} \mu_\phi([a_0^{n-1}]) \mu_\phi \left\{ \underline{y} : \tau_{[a_0^{n-1}]}(\underline{y}) \mu_\phi([a_0^{n-1}]) > e^{nt/2} \right\} \\ &\leq C' e^{-c' e^{nt/2}} \end{aligned}$$

for some  $c', C' > 0$ .

Since the bound for  $T_1$  is (much) smaller than the bound for  $T_2$ , we can bound  $T_1 + T_2$  by a constant times  $e^{-Bnt^2}$ . This yields (11).

We now turn to the proof of (12). We have

$$\begin{aligned}
& (\mu_\phi \otimes \mu_\phi) \left\{ (\underline{x}, \underline{y}) : \frac{1}{n} \log W_n(\underline{x}, \underline{y}) < h(\mu_\phi) - t \right\} \\
&= (\mu_\phi \otimes \mu_\phi) \left\{ (\underline{x}, \underline{y}) : -\frac{1}{n} \log W_n(\underline{x}, \underline{y}) - \frac{1}{n} \log \mu_\phi([x_0^{n-1}]) + \frac{1}{n} \log \mu_\phi([x_0^{n-1}]) + h(\mu_\phi) > t \right\} \\
&\leq (\mu_\phi \otimes \mu_\phi) \left\{ (\underline{x}, \underline{y}) : -\frac{1}{n} \log [W_n(\underline{x}, \underline{y}) \mu_\phi([x_0^{n-1}])] > \frac{t}{2} \right\} \\
&\quad + \mu_\phi \left\{ \underline{x} : \frac{1}{n} \log \mu_\phi([x_0^{n-1}]) + h(\mu_\phi) > \frac{t}{2} \right\} \\
&= T'_1 + T'_2.
\end{aligned}$$

Proceeding as for  $T_2$  (applying Corollary 3.3 to  $f = \phi$ ) we obtain the upper bound

$$\begin{aligned}
T'_2 &\leq \mu_\phi \left\{ \frac{1}{n} (\phi + \dots + \phi \circ \sigma^{n-1}) - \int \phi d\mu_\phi > \frac{t}{2} - \frac{1}{n} \log C \right\} \\
&\leq e^{-Bnt^2}
\end{aligned}$$

for some  $C''' > 0$  and for every  $t > 2 \log C$ .

To bound  $T'_1$  we use the following lemma (Lemma 9 in [1]).

**Lemma 4.3** ([1]). *For any  $v > 0$  and for any  $a_0^{n-1}$  such that  $v\mu_\phi([a_0^{n-1}]) \leq 1/2$ , one has*

$$\lambda_1 \leq -\frac{\log \mu_\phi \left\{ \tau_{[a_0^{n-1}]} > v \right\}}{v\mu_\phi([a_0^{n-1}])} \leq \lambda_2,$$

where  $\lambda_1, \lambda_2$  are the constants appearing in Lemma 4.2.

The previous lemma implies that

$$\mu_\phi \left\{ \tau_{[a_0^{n-1}]} \mu_\phi([a_0^{n-1}]) < v \right\} \leq 1 - e^{-v\lambda_2} \leq \lambda_2 v$$

provided that  $v\mu_\phi([a_0^{n-1}]) \leq 1/2$ . Taking  $v = e^{-nt/2}$  it follows that

$$\begin{aligned}
T'_1 &= \sum_{a_0^{n-1}} \mu_\phi([a_0^{n-1}]) \mu_\phi \left\{ \underline{y} : \tau_{[a_0^{n-1}]}(\underline{y}) \mu_\phi([a_0^{n-1}]) < e^{-nt/2} \right\} \\
&\leq \lambda_2 e^{-nt/2}.
\end{aligned}$$

This inequality holds if  $e^{-nt/2} \mu_\phi([a_0^{n-1}]) \leq 1/2$ , which is the case for any  $n \geq 1$  if  $t \geq 2 \log 2$ .

Inequality (12) follows from the bound we get for  $T'_1 + T'_2$ . But the bound for  $T'_1$  is bigger than the one for  $T'_2$ , whence the result.

The proof of the theorem is complete.  $\square$

## APPENDIX A. PROOF OF LEMMA 4.1

We start with the following identity:

$$\begin{aligned}
 \hat{h}_k(x_0^{n-1}) &= - \sum_{a_0^{k-1} \in A^k} \mathcal{E}_k(a_0^{k-1}; x_0^{n-1}) \log \frac{\mathcal{E}_k(a_0^{k-1}; x_0^{n-1})}{\mathcal{E}_{k-1}(a_0^{k-2}; x_0^{n-1})} \\
 (13) \quad &= - \sum_{a_0^{k-1} \in A^k} \mathcal{E}_k(a_0^{k-1}; x_0^{n-1}) \log \frac{\mu_\phi([a_0^{k-1}])}{\mu_\phi([a_1^{k-1}])} + \hat{\Delta}_k(x_0^{n-1}),
 \end{aligned}$$

where

$$\begin{aligned}
 \hat{\Delta}_k(x_0^{n-1}) &:= \\
 &- \sum_{a_0^{k-1} \in A^k} \mathcal{E}_k(a_0^{k-1}; x_0^{n-1}) \log \frac{\mathcal{E}_k(a_0^{k-1}; x_0^{n-1})}{\mathcal{E}_{k-1}(a_0^{k-2}; x_0^{n-1})} + \sum_{a_0^{k-1} \in A^k} \mathcal{E}_k(a_0^{k-1}; x_0^{n-1}) \log \frac{\mu_\phi([a_0^{k-1}])}{\mu_\phi([a_1^{k-1}])} = \\
 &- \sum_{a_0^{k-1} \in A^k} \mathcal{E}_k(a_0^{k-1}; x_0^{n-1}) \log \frac{\mathcal{E}_k(a_0^{k-1}; x_0^{n-1})}{\mu_\phi([a_0^{k-1}])} + \sum_{a_0^{k-1} \in A^k} \mathcal{E}_k(a_0^{k-1}; x_0^{n-1}) \log \frac{\mathcal{E}_{k-1}(a_0^{k-2}; x_0^{n-1})}{\mu_\phi([a_1^{k-1}])} = \\
 (14) \quad &- H_k(\mathcal{E}_k(\cdot; x_0^{n-1}) | \mu_\phi) + H_{k-1}(\mathcal{E}_{k-1}(\cdot; x_0^{n-1}) | \mu_\phi),
 \end{aligned}$$

where

$$H_k(\eta | \mu_\phi) = \sum_{a_0^{k-1} \in A^k} \eta([a_0^{k-1}]) \log \frac{\eta([a_0^{k-1}])}{\mu_\phi([a_0^{k-1}])}$$

is the  $k$ -block relative entropy of  $\eta$  with respect to  $\mu_\phi$ . The second term in (14) is equal to  $H_{k-1}(\mathcal{E}_{k-1}(\cdot; x_0^{n-1}) | \mu_\phi)$  because of the following two facts. First,  $\sum_{a_0 \in A} \mathcal{E}_k(a_0^{k-1}; x_0^{n-1}) = \mathcal{E}_{k-1}(a_1^{k-1}; x_0^{n-1})$ . This is because  $\mathcal{E}_k(\cdot; x_0^{n-1})$  is a locally shift-invariant probability measure on  $A^k$ . Second,  $\sum_{a_{k-1} \in A} \mathcal{E}_k(a_0^{k-1}; x_0^{n-1}) = \mathcal{E}_{k-1}(a_0^{k-2}; x_0^{n-1})$ , because the family  $(\mathcal{E}_k(\cdot; x_0^{n-1}))_{k=1,2,\dots}$  is consistent.

The quantity  $|\hat{\Delta}_k(x_0^{n-1})|$  is bounded above by  $(M|A|^k)/n$  according to [10, formula (4.16)], where  $M > 0$  is a constant.

Now we deal with the first term in (13). We first introduce the function

$$\phi_k(\underline{y}) := \log \frac{\mu_\phi([y_0^{k-1}])}{\mu_\phi([y_1^{k-1}])}$$

which is a locally constant function on cylinders of length  $k$ . It is easy to verify that  $\|\phi - \phi_k\|_\infty \leq |\phi|_\theta \theta^k$  (this follows at once from [12, Prop. 3.2 p. 37]). We get that

$$- \sum_{a_0^{k-1} \in A^k} \mathcal{E}_k(a_0^{k-1}; x_0^{n-1}) \log \frac{\mu_\phi([a_0^{k-1}])}{\mu_\phi([a_1^{k-1}])} = \frac{1}{n} \sum_{j=0}^{n-1} (-\phi(\sigma^j \underline{x})) + \mathcal{O}(\theta^k).$$

The proof of the lemma is complete.

## REFERENCES

- [1] M. Abadi. Sharp error terms and necessary conditions for exponential hitting times in mixing processes. *Ann. Probab.*, 32(1A):243–264, 2004.
- [2] A. Antos and I. Kontoyiannis. Convergence properties of functional estimates for discrete distributions. *Random Structures Algorithms*, 19(3-4):163–193, 2001. Analysis of algorithms (Krynica Morska, 2000).
- [3] R. Bowen. *Equilibrium states and the ergodic theory of Anosov diffeomorphisms*, volume 470 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, revised edition, 2008.
- [4] J.-R. Chazottes, P. Collet, and B. Schmitt. Statistical consequences of the Devroye inequality for processes. Applications to a class of non-uniformly hyperbolic dynamical systems. *Nonlinearity*, 18(5):2341–2364, 2005.
- [5] J.-R. Chazottes and D. Gabrielli. Large deviations for empirical entropies of  $g$ -measures. *Nonlinearity*, 18(6):2545–2563, 2005.
- [6] J.-R. Chazottes and E. Ugalde. Entropy estimation and fluctuations of hitting and recurrence times for Gibbsian sources. *Discrete Contin. Dyn. Syst. Ser. B*, 5(3):565–586, 2005.
- [7] P. Collet, A. Galves, and B. Schmitt. Repetition times for Gibbsian sources. *Nonlinearity*, 12(4):1225–1237, 1999.
- [8] P. Collet, S. Martínez, and B. Schmitt. Exponential inequalities for dynamical measures of expanding maps of the interval. *Probab. Theory Related Fields*, 123(3):301–322, 2002.
- [9] D. P. Dubhashi and A. Panconesi. *Concentration of measure for the analysis of randomized algorithms*. Cambridge University Press, Cambridge, 2009.
- [10] D. Gabrielli, A. Galves, and D. Guiol. Fluctuations of the empirical entropies of a chain of infinite order. *Math. Phys. Electron. J.*, 9:Paper 5, 17 pp. (electronic), 2003.
- [11] M. Ledoux. *The concentration of measure phenomenon*, volume 89 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 2001.
- [12] W. Parry and M. Pollicott. Zeta functions and the periodic orbit structure of hyperbolic dynamics. *Astérisque*, (187-188):268, 1990.
- [13] P. C. Shields. *The ergodic theory of discrete sample paths*, volume 13 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 1996.

CPHT, CNRS-ÉCOLE POLYTECHNIQUE, 91128 PALAISEAU CEDEX, FRANCE

EMAIL ADDRESS: [jeanrene@cpht.polytechnique.fr](mailto:jeanrene@cpht.polytechnique.fr)

EMAIL ADDRESS: [maldonado@cpht.polytechnique.fr](mailto:maldonado@cpht.polytechnique.fr)